A/B Testing

Author: Palak Sarawagi

Causal Effect:

The causal effect quantifies the change in X, which caused a change in Y.

For example:

To determine if a new drug is effective for headaches. The causal effect would be the difference between the headache when the person takes the drug, and the headache when the person does not take the drug.

Let us look at the following two scenarios. In the example below, the drug is the treatment variable (X), and headache is the outcome variable (Y) for Person A.

In Scenario 1:

Person A is treated with the drug (X=1)

Person A's headache, when treated with the drug (X=1), can be written as:

$Y_a(1)$ or $Y_a \mid X_a = 1$

In Scenario 2:

Person A is not treated with the drug (X=0)

Person A's headache, when not treated with the drug (X=0) can be written as:

$Y_a(0)$ or $Y_a \mid X_a = 0$

The causal effect for the above example can be written and explained as follows:

$Y_a(1) - Y_a(0)$, i.e., The difference between the outcome variable Y when the person took the drug and the outcome variable Y when the person did not.

However, the fundamental problem with calculating causal effects is that only one world exists. Person A cannot be - not treated with the drug and be treated with the drug at the same time. Therefore, it is not possible to calculate the causal effect as mentioned above.

Counterfactual

In simple words, counterfactual is something that has not happened. Let us understand counterfactual, in the context of our above example:

As mentioned above, Person A cannot be - not treated with the drug and be treated with the drug at the same time.

Let us say that Person A is in Scenario 1, where they are treated with a drug Y(1). In this case, Scenario 2, where they are not treated with the drug, cannot exist Y(0). We can say that the missing scenario (or scenario that did not happen), Scenario 2, is the counterfactual.

Counterfactuals are why causal effects cannot be treated for an individual.

Selection

Let us try to understand Selection in continuation of our previous example. So far, we have understood that causal effects are difficult to calculate because of counterfactuals.
How do we resolve this problem?

One may say that we can have Person A in Scenario 1, where they are treated with a drug $Y_a(1)$, and have Person B in Scenario 2, where they are not treated with the drug $Y_b(0)$. The causal effect can be calculated as $Y_a(1) - Y_b(0)$.

But, the above equation will give us an incorrect causal effect because the two people in the equation are different. Before the treatment, Person A may have a more severe headache as compared to Person B and vice-versa; thus the causal effect we calculate will be incorrect.

Selection is the measure of how two subjects are different from each other. To calculate the correct causal effect of the treatment, the selection between two people, A and B, needs to be zero.

Randomized Experiment

We usually perform a randomized experiment to balance and nullify the selection to ensure no selection. In a randomized experiment, people are randomly chosen to be experimented on. Furthermore, if the person will receive the treatment (treated group) or not receive the treatment (control group) is also determined randomly. The random selection balances the selection in both groups.

But how do we make this random selection?
We take a significant number of people and unbiasedly put them in two groups (treated or controlled). For example: In our above example, if we need to put 500 people in the two groups.
We can flip a coin for each person:
If the person gets heads, we put them in the treated group, i.e., treat them with drugs.
If the person gets tails, we put them in the controlled group, i.e., do not treat them with drugs.

The randomized experiment is a great way to have zero selection. The causal effect can then be determined as follows:
$\{E[Y_i(1)]-E[Y_j(0)]\}$

Here,
$E[Y_i(1)]$ = Expected outcome (mean) of all the people who received the treatment
and
$E[Y_j(0)]$ = Expected outcome (mean) of all the people who did not receive the treatment

Correlation v/s Causation

Although correlation and causation are two related ideas, they have different meanings.

Correlation is used to understand the relationship between two variables. However, it does not imply that one variable causes a change in another variable.

On contrary, causation means that a change in one variable causes a change in another variable.

An understanding of the difference between correlation and causation is very important. If not understood carefully, it can lead to a lot of misunderstandings and wrong inferences. The same can be illustrated using the following examples:

Example 1:

Some people are protesting that sales of selfie sticks increase crime. Clearly, logic isn't the strongest here. This notion can be explained as:

The increase in the sale of selfie sticks is caused due to an increase in the population. Similarly, the increase in crime rate is caused due to an increase in population. Here, the population was an unobservable variable that caused the increase in crime rates and the sale of selfie sticks. Clearly, the crime rate and the sale of selfie sticks are correlated (both increase with time) but do not cause each other.

Example 2:

Some people are superstitious and believe that a black cat crossing your path causes bad luck. This can be explained as:

In earlier days, when a cat used to run across a path, usually, there used to be a larger animal chasing it as prey. This used to put humans under the threat of larger animals (bad luck). Here the black cat and bad luck can be correlated but do not cause each other. It was the presence of a bigger animal that caused the bad luck.

A/B Testing Example:

Testing if the new version of the website improves user engagement

To test if the new version of the website improves user engagement for a website, a Randomized Control Trial (RCT) should be used.

Why do we perform RCT?

To do an observational study for two versions (v0 and v1) of the website, we need to eliminate the bias or differences between the two groups of people participating in the study. So, we take a large sample of people and assign the people randomly to a group (v0 or v1). This random allocation distributes the two groups' influential factors or differences equally. Thus, indicating zero selection.

How to perform an RCT?

To perform the Randomized Control Trial, the following steps should be performed:

(i) **Determining the test:** To test if v1 of the website has better user engagement than v0. We will use the metric pageview and time on page to determine the outcome of the test. An increase in pageviews and time on the page suggests better user engagement.

(ii) **Defining the data to be tested:** To perform this test, we will choose users who have been using the website only for the past six months. (If the users have been using the website for a more extended period, they will be comfortable with the previous UI because of practice)

(iii) **Randomly allocating:** The selected people from the above step will be randomly allocated to two groups (Group A will use v0, Group B will use v1). This simple randomization can be performed easily using a process as simple as a flip of a fair coin or using a tool.

(iv) **Conducting the test:** The two groups of people determined from the above step are then studied over a period, till sufficient data is generated (usually 8-12 weeks in such cases) to determine a pattern in the outcome. In the above case, group A which continues to use the v0 is the control group, and group B which uses the new v1 is the treatment group.

(v) **Determining and measuring the outcome:** Once we have collected enough data required by our experimental design, we test our null hypothesis (the new v1 increases user engagement). We can run a regression to see if the new version has a statistically significant impact on page time.

Note: We need to be cognizant of the fact that RCTs may incur Type 1 or Type II errors. We may get a result that the null hypothesis is true when it is actually not true (false positive) or we may get a result that the null hypothesis is false when it is actually true (false negative)

Other Limitations:

(i) Page time is not always the best metric to calculate the user engagement.

(ii) It would be difficult to determine if only a part of the new version helps better user engagement than the other.